DATA SUITABILITY REVIEW PROCESS

Center for Environmental Information and Statistics

center for environmental information and statistics

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The **Center for Environmental Information and Statistics (CEIS)** is conducting an ongoing review of EPA's major databases. As EPA's databases become publicly available, they are being employed for uses other than the primary use for which they were originally designed. The CEIS review is focused on the suitability of EPA's databases for these secondary uses.

The 31 major databases that are being reviewed include those that are most heavily relied on inside and outside of EPA for environmental information. The review process results in a Major Findings Document for each database. The major findings of each database are derived from available documentation, a descriptive profile developed with the assistance of the Program Office that maintains the database, a statistical profile developed by CEIS, and case studies conducted by CEIS.

The Major Findings Documents will be reviewed by the relevant Program Offices. CEIS will work closely with the Program Offices at every stage and will incorporate their input in the Document. Subsequently, the Major Findings Documents will be circulated for peer review.

The CEIS review of data suitability addresses issues surrounding the appropriateness of the databases for secondary uses. At the same time, another CEIS initiative, the Customer Survey, addresses customers' needs for and uses of environmental information. Assessing the two together will provide a clearer picture of what EPA has and what is being asked of EPA. Information from both the review of data suitability and the customers' requirements will allow a Gaps Analysis, the basis for CEIS to develop a Data Enhancement Plan. As EPA databases are enhanced, the CEIS will continue to reevaluate and conduct data suitability reviews as required.

I. Introduction

The **United States Environmental Protection Agency (EPA**), in its regulatory and policy setting role, collects environmental data. In general, each of EPA's databases serves as a repository for the data collected to meet a primary objective such as collection of data in order to ensure compliance with a particular regulation, meeting permitting requirements, meeting right-to-know requirements, etc. Increasingly, EPA's databases are becoming publicly available and a number of secondary uses of these databases are emerging. These secondary uses include evaluating the local state of the environment, identifying pollution sources and hot spots, promoting environmental education, assessing corporate accountability, etc.

The Center for Environmental Information and Statistics (CEIS) plays a crucial role in reviewing the databases in terms of their suitability for secondary uses. At a minimum, it is important that users of a database understand what is in the database, its geographical as well as temporal coverage, and its limitations, prior to using it for purposes other than what it was designed for. This paper discusses the ongoing CEIS process for reviewing the major EPA databases and their applicability for secondary uses.

The current review focuses on 31 major databases that are most heavily relied on inside and outside of EPA for information on the current state of the environment, on changes over time, and on the nature of the specific actions being taken by EPA and its counterpart agencies at the state and local level to protect and improve environmental quality. The 31 major databases being reviewed are:

OFFICE OF WATER

- Storage and Retrieval of Water Quality Information (STORET X)
- Safe Drinking Water Information System (SDWIS)
- Water Body System
- Reach File (RF3) and the National Hydrography Dataset (NHD)
- Ocean Data Evaluation System (ODES)
- Clean Water Needs Survey (NEEDS)
- Environmental Monitoring Methods Inventory (EMMI)
- Information Collection Rule

OFFICE OF AIR

- Aerometric Information Retrieval System - Air Quality Subsystem (AIRS-AQS)
- Emissions Tracking Subsystem (ETS) - Acid Rain Program
- Aerometric Information Retrieval System - AIRS Emission Subsystem (AIRS-AES)
- Title VI Allowance Tracking System
- Environmental Radiation Ambient Monitoring System (ERAMS)
- Findings and Required Elements Data System (FREDS)
- Sample Tracking and Data Management System (STDMS)

OFFICE OF SOLID WASTE AND EMERGENCY RESPONSE

- Resource Conservation and Recovery Information System (RCRIS)
- Comprehensive Environmental Response, Compensation, and Liability Information System (CERCLIS)
- Biennial Reporting System (BRS)

OFFICE OF POLLUTION PREVENTION AND TOXIC SUBSTANCES

- Toxics Release Inventory (TRI)
- National Pesticide Information Retrieval System (NPIRS)
- Chemicals in Commerce Information System (CICIS)

OFFICE OF ENFORCEMENT AND COMPLIANCE ASSURANCE

- Permit Compliance System (PCS)
- Aerometric Information Retrieval System - AIRS Facility Subsystem (AIRS-AFS)
- Compliance Subsystem of CERCLIS
- Compliance Subsystem of RCRIS
- National Compliance Database (NCDB)
- Waste International Tracking System (WITS)
- Enforcement Docket (DCK)
- National Asbestos Registry System (NARS)
- Section Seven Tracking System (SSTS)
- Compliance Subsystem of SDWIS

The CEIS review of these databases addresses the following two questions.

1. What is the current quality of data in the database?

2. What is the potential for cross database analysis, i.e., for integrating data from that database with similar data from other EPA databases?

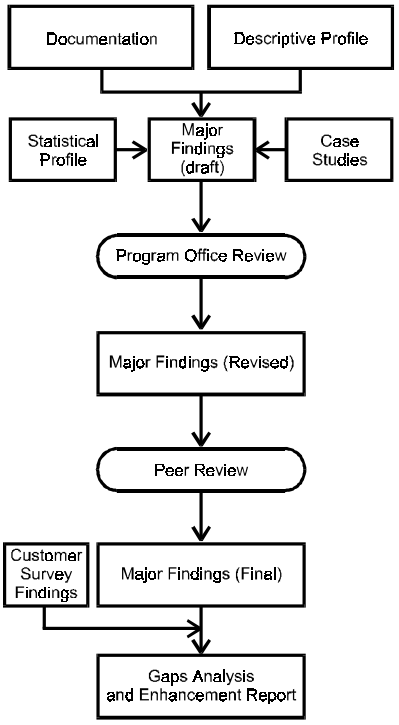
1. Quality

For this review the quality of the database is a function of the specific purpose to which a given user is considering applying it. Quality is therefore a function of the intended use. It is important to understand that Program Offices assure the quality of the data in the databases they are responsible for. However, CEIS, in its review, addresses quality as the applicability of the data for potential secondary uses. For this reason, the review process is assembling information about each of the databases to determine the suitability of the data for meeting secondary uses.

2. Cross Database Analysis

Information from any two databases containing similar environmental data may be jointly analyzed at some scale (e.g., by aggregating the data in each database to the national level for an entire year). The current review focuses on the spatial characteristics, temporal features, and the EPA identification numbers associated with the data in each of the 31 databases that would facilitate analysis involving two or more databases.

Data Suitability Review



This is a continuous process with the results of the Gaps Analysis and Enhancement Report leading to database enhancements and restarting the Database Suitability Reviews.

2. Activities Associated with the Data Suitability Reviews

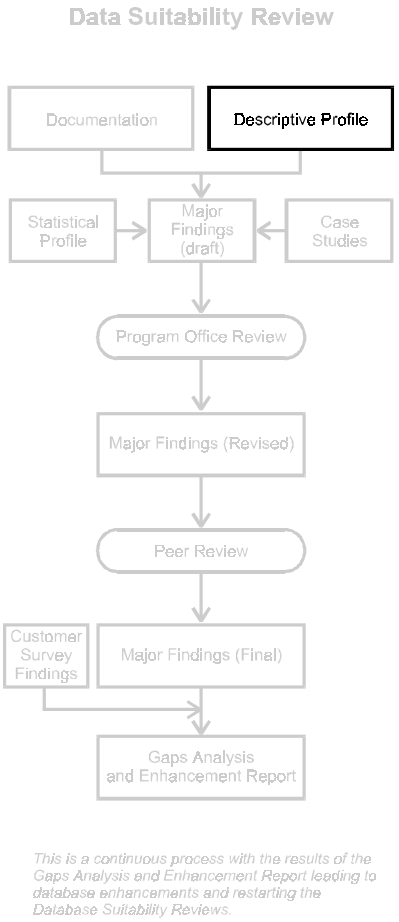
The CEIS data suitability review process has many components (as outlined in the *Data Suitability Review* figure). CEIS works closely with the Program Offices that are responsible for maintaining the databases. Program Office participation is crucial in obtaining documentation and a descriptive profile, and, more importantly, in critiquing and validating the findings of the CEIS review process. The findings of the review process will also be subject to peer review. The findings from the Data Suitability Reviews of the databases will then be assessed against the findings from the Customer Survey, another key activity of the CEIS. The former addresses the supply side of environmental information and the latter addresses the demand side. Examining the two sides together will assist the CEIS in performing a Gaps Analysis as well as developing a strategy for Database Enhancement. The activities outlined in the figure are described in the following sections.

Data Suitability Review

This is a continuous process with the results of the Gaps Analysis and Enhancement Report leading to database enhancements and restarting the Database Suitability Reviews.

a. Review of Database Documentation

The first step in reviewing a database is obtaining the available relevant documentation on the database, both from the Program Office and any additional sources. The documentation is used to understand the details of the database, including the types of data collected, consistency checks performed, and other activities associated with data collection and quality assurance. The documentation also indicates the extent and means of information dissemination. It is critical to obtain the information about the data (also known as metadata) as it is important that users be able to refer to these sources for details about the data. Secondary sources (other than the Program Offices) include Envirofacts and websites of user groups such as RTKNet. User groups often highlight the limitations that they encounter while using the database for secondary uses.



b. Descriptive Profile

A descriptive profile of each database is developed from a questionnaire that serves as the basic instrument for capturing and recording information about each database. This questionnaire is completed by the Program Office that maintains the database. An example of a current Database Questionnaire appears in Appendix A. Each Database Questionnaire consists of approximately 70 questions grouped into six sections. Each section addresses an aspect or set of characteristics in order to review the suitability of the database for secondary uses. The six sections are:

1. Purpose for which the database has been established
2. Comprehensiveness
3. Spatial characteristics
4. Temporal characteristics
5. Internal consistency
6. Comparability with and ability to integrate with other databases

There are currently two different versions of the Database Questionnaire for the following two types of data systems.

- Monitoring data systems
- Facility data systems

c. Statistical Profile

The statistical profile calls for an actual hands-on statistical assessment of all or part of the database being reviewed. This hands-on statistical assessment is essential to ensure the preparation of an accurate appraisal and a description of the real data contained in the database and of any consequent limitations that must be placed on the interpretation of summary statistics prepared from these data.

The hands-on statistical assessment therefore goes beyond merely describing characteristics of the data (e.g., spatial identifiers), and characteristics of the data collection activity (e.g., the extent to which random sampling was utilized). In addition, it deals with such issues as the extent of missing data and the distribution of the data over key variables that might be used for secondary analyses.

Appendix B provides the broad categories for the statistical assessment. The specified statistical analysis routines vary with the type of database and the type of variable. The protocol for statistical assessment will continue to undergo further revision and refinement as further experience is gained applying the protocol to various kinds of databases.

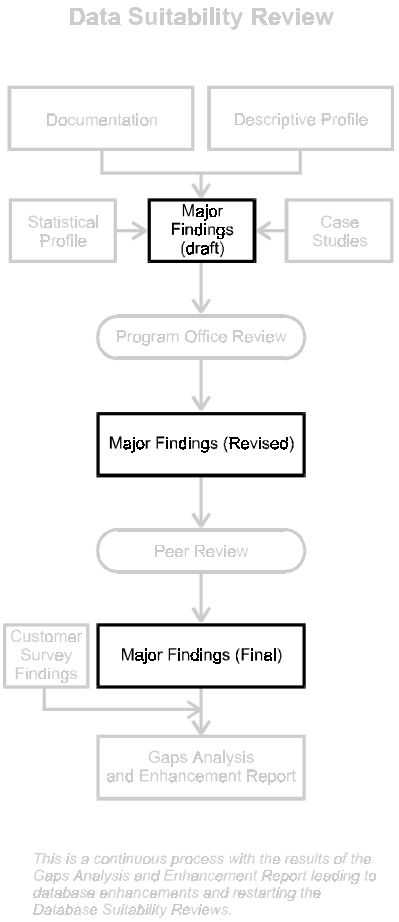
d. Case Studies

An integral part of the evaluation of the suitability of databases for secondary uses is to actually try to use a database in this manner. Clearly the ability of a database to handle any one particular situation is by no means proof of its general robustness to handle all situations. However, the effort to tease out data for a specific application provides insight into the database in general.

Two levels of case studies were considered for the databases. The first is a hypothetical, yet plausible, application. The second is based on some desired information as determined by CEIS surveys of users' needs. In the first instance, CEIS tried to devise a situation in which a given database would be a logical source of information for a hypothetical question. For instance, one of these questions assumed that chicken farmers in Maryland along the Pocomoke River thought the *pfiesteria* problem of late 1997 could have been caused by something other than chicken manure, potentially other sources of contaminants going into the river. This led to examination of the Permits Compliance System (PCS) to see if there were violating discharges into the river during this time. The objective was not to solve the *pfiesteria* problem, but rather to investigate the ability to use PCS to analyze the situation.

The second level of case studies reviews problems that are being posed by the interested public, as viewed through CEIS survey efforts. At this time, 'twenty questions' (Appendix C) have been assembled, based on input from customer surveys conducted by the CEIS. Future case studies will be oriented to see whether and which EPA databases can answer these concerns.

The case studies are particularly helpful in bringing to light not just nuances of the databases, but also information about the ease with which an interested user can access the databases. Several EPA databases are available in various formats from multiple sources. These sources include the originating Program Office's master database, Envirofacts, RTKNet versions, and CD-ROM versions distributed by private vendors. Since each of these versions has its own attributes, these must be considered along with the operation of the database itself. Various runs of the case studies should shed light on the comparative features of the range of publicly available data.



e. Document Major Findings

The information gathered from the documentation and the descriptive profile will be used initially to document the major findings about the database.

In order to present the review in a user-friendly manner, the information will be organized in a uniform format for all databases. After an initial summary description of the database, all relevant information will be presented as answers to nine key questions that a user is likely to ask in reviewing the database to assess its suitability.

In order to present the review in a user-friendly manner, the CEIS plans to organize the information in a uniform format for all databases. After an initial summary description of the database, all relevant information will be presented as answers to nine key questions that a user is likely to ask in reviewing the database to assess its suitability.

1. How comprehensive is the database?
Identifies the kinds of information contained in the database (e.g. types of pollutants, facilities, permits, and acquisition methods).

2. Can the database be used for spatial analysis?
Indicates whether the data are available at each of various geographical scales such as ZIP code, latitude and longitude, county code, State code, etc.

3. Can the database be used for temporal analysis?
Shows if the data are collected on a fixed schedule (daily, monthly, yearly, etc.) and if the data are sufficiently consistent over time to allow period-to-period comparisons.

4. How consistent are the variables over space and time?
Indicates the degree of internal consistency allowing comparisons across space (facilities, Regions, etc.) and over time (monthly, yearly, etc.)

5. Can data be linked with information from other databases?
Provides information that can be used to determine if data can be linked with other databases based on common characteristics such as facility identification numbers, latitude and longitude, geographical codes, etc.

6. How accurate are the data?

Information from data quality checks performed by the Program Office as well as from statistical analysis performed by the CEIS appears here.

7. What are the limitations?

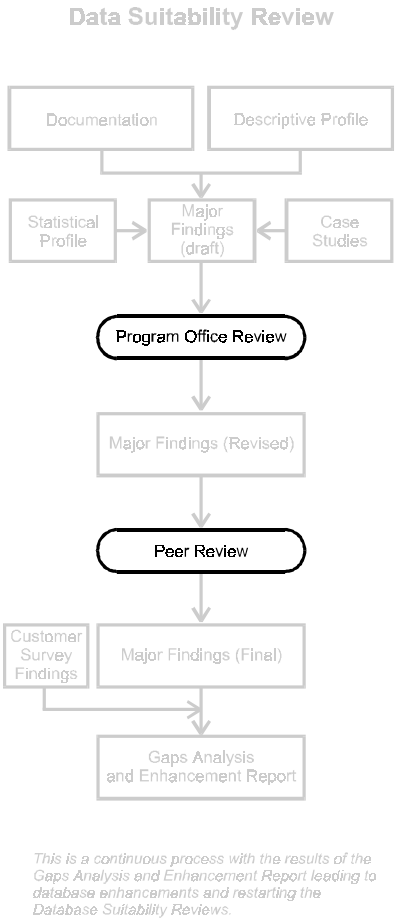
Each of EPA's databases has a primary purpose for which it was developed and is maintained. As the databases are reviewed for suitability for alternate uses, it is important to understand the constraints and limitations of the database.

8. How can I get information?

Identifies the formats in which the database is available such as printed form, diskettes, CD-ROM, online access etc., along with names, addresses, and phone numbers to contact for detailed information.

9. Is there documentation?

A quality database requires documentation to support it, such as information on data collection methods, quality assurance mechanisms, data management, users' guides, and information dissemination. Details on the availability of such documentation appears here.



f. Program Office Review and Peer Review

The findings from the case studies and the statistical profiles will be incorporated into the major findings as the analyses proceed. The Draft Major Findings Documents will be subject to review by the Program Offices and revised accordingly. The Revised Major Findings Document will then be peer reviewed and those comments incorporated to produce the Final Major Findings Document.

g. Customer Survey Findings

To better understand the needs of their customers, the EPA's Center for Environmental Information and Statistics (CEIS) and the Environmental Monitoring for Public Access and Community Tracking (EMPACT) program are conducting a comprehensive customer survey. Although the research is being conducted principally for, and on behalf of, CEIS and EMPACT, the results are applicable to all EPA programs, as well as to a broad range of Federal, State, local, non-governmental organizations, and information users and providers.

The CEIS has been conducting facilitated public meetings focusing on needs, preferences, and behavior with respect to the use of environmentally related data and information since July 1997. The meetings include a cross section of EPA customers and stakeholders, including:

- Industry Associations
- Public Health Officials
- Environmental Organizations
- Environmental Justice Groups
- Community Organizations and Community Members
- Educators and School Administrators
- Researchers
- State and Local Government Representatives
- Community Right-to-Know Groups

The focus is on how the customers or stakeholders themselves describe the processes of identifying, obtaining, interpreting, and using environmentally-related data or information in their professional or work activities and also in their personal lives. The group meeting process also produces insight into the views of customers on their needs for environmental information, strengths and weaknesses of various specific datasets, internet sites, and other topics related to the use of information.

h. Gaps Analysis and Enhancement Report

Findings from the Data Suitability Review and the Customer Survey can be analyzed in conjunction with each other to identify and characterize gaps in demand for and supply of environmental information. It is likely that several issues will arise that may require a range of different enhancement options. While some issues that arise may be simple to address, others may be extremely complex. For example, there are several versions of a database that may be electronically available to the public. Sometimes it is possible to access a database by visiting the Program Office web site, via EPA's Envirofacts, potentially through CEIS, or through RTKNet. There may be a simple solution of designating one site as the official EPA site to access the database. A more complex issue may be that customers indicate a need to understand the local state of the environment for a particular area where the use of the data may require assumptions and complex calculations. It is important to identify the issues that arise from the Gaps Analysis; to design an Enhancement Plan to meet both EPA's and customers' needs; and to ensure that EPA's data are being used appropriately in these secondary uses.

APPENDIX A

Database Questionnaire

Assessment of the Potential Suitability of
The AIRS - AQS Database For Various Uses

AREAS OF DATA QUALITY CONCERN

Please note that the following questions are written for databases that contain ambient environmental monitoring data. Parallel sets of questions will be prepared for each other type of database containing environmental information, including databases containing: emission/discharge/release data, compliance data, ambient data from the investigation of specific sites, or administrative records data relevant for assessing the state of the environment.

QUESTIONS TO BE ASKED FOR DATABASES CONTAINING AMBIENT ENVIRONMENTAL DATA

1. Purpose

The questions below seek to answer the following broad question:

"For what purposes is this data collection activity being conducted?"

a. What is the purpose of this data collection activity?

(1) What was the primary purpose for which this data collection activity was originally designed? - **To evaluate as to whether or not areas of the country were/are meeting the National Ambient Air Quality Standards for criteria pollutants which were established to protect the health of Americans**

(2) What were the secondary purposes that those who initially established this data collection activity had in mind, if any? - **To collect and analyze ambient air quality data for research purposes and evaluation of trends in the quality of air in America**

(3) What additional secondary purposes have become established since then that are considered significant by those now managing or conducting this data collection activity? - **AIRS currently collects information on pollutants that form ozone. The reduction of these precursor emissions is an essential part in reducing the concentrations of ozone in the atmosphere. Additionally, we store data from various agencies throughout the world (about 50 countries). AQS also stores information on trace metals...**

2. Comprehensiveness: What is measured at each site?

"What chemicals or other aspects of environmental quality does the data in this database characterize?"

a. What are the chemicals or other parameters for which data was obtained at each data collection site? **AIRS is able to store nearly 3,000 different pollutants. If you need a complete list, one can be provided.**

b. How is each measured variable or parameter specified? (E.g., each chemical to be measured is specified by a common name plus its CAS number) **The parameter is specified by supplying a parameter code that has been created by the AIRS system administrator. In addition, we store the compound name, CAS number (if available), and an abbreviation.**

3. Spatial Characteristics

The specific questions below seek to answer the following broad question:

"How were the sites at which data collection takes place selected? How are these locations specified? With what locational accuracy?"

a. Overall design of the network

(1) What is the geographic universe of concern? - **United States and 50 Countries**
(2) What is the smallest meaningful sampling unit? **- a single monitor**
(3) How is the universe stratified? - **The data is stratified by sites, counties, and states**

b. What specific procedures were used to select the data collection sites used?

(1) What design was used to determine the general location of each sample collection site (e.g., "locate one site on the Lower Mississippi between Baton Rouge and New Orleans") Possible procedures: statistical selection, directed selection (e.g., "have one site in each county in the State"), locate at "hot spots," locate at population centers. **- There are various reasons for why monitors were placed where they were. For specifics for monitors composing the State and Local Air Monitoring Stations (SLAMS), Photochemical Assessment Monitoring Stations (PAMS), and National Air Monitoring Stations (NAMS), siting criteria can be found in 40CFR Part 58, Appendix D. Monitors have also been sited for special purpose monitoring by the USEPA, state, or local agency. These reasons include: "hot spot" monitoring, population exposure, background monitoring, source oriented monitoring or location for maximum concentration. Additional design considerations may be taken into account to meet more stringent State standards.**

(2) What procedures were used to determine the specific data collection point at each site (e.g., the site mentioned above on the Lower Mississippi River between Baton Rouge and New Orleans, how was the specific location of the data collection point, in river miles from the mouth of the Mississippi, selected? How was the distance from shore at which the physical sample was to be taken selected? How was the water depth at which the physical sample was to taken selected?") Possible procedures: statistical selection, expert judgment ("where possible, do not collect a physical sample at a point closer than ten feet from the river bank"), or convenience ("collect from the nearest bridge at the midpoint"). - **Again, the answers will vary depending on the purposes of the collecting agency.**

c. Spatial Coordinates Used and Locational Accuracy

(1) How is the location of each site described/specified within the database? (i.e., by lat/long, by UTM, by a physical description of the location [i.e., from a point in Mill Creek 50 feet north of the south footing of the Highway 42 bridge near Jones City, Tennessee] - **lat/long and UTM**

(2) What is the degree of locational accuracy of the description of the site location contained in the data base? (e.g. within 100 feet, within 15 feet, within one second of latitude and 4 seconds of longitude) - **Lat/long coordinates can be entered to an accuracy of 1/10,000 of a second. UTM's can be reported to within 1 meter**

4. Temporal Characteristics

"What is the frequency with which data collection occurs at each site? How long does it take for newly collected data to become available on the database? Are there any significant variations in the data values over time (e.g., seasonally)?"

a. What is the frequency of sample collection? - **The answers will vary. Typically, for the gaseous pollutants, continuous samples are common. For PM10 data, every 6th day is common. For other parameters, the answers vary tremendously.**

b. Is there documentation available that tells:

(1) On what date did this data collection activity begin? - **Yes**
(2) On what dates were specific data collection sites added or discontinued? - **Yes**

c. Are changes over time in the measurement and collection process documented for:

(1) For the overall conduct of this data collection effort? - **Yes** (on a per reporting organization basis)
(2) For the measurement methodologies used? - **Yes**

d. Is there documentation of any seasonal or diurnal factors of concern in the conduct of this data collection effort? - **Yes**

e. Is there documentation of any changes over time in any external factors relevant to this data collection effort? - **No**

f. Timeliness -

(1) What is the elapsed time between when a physical sample is collected in the field (i.e., at a data collection site) and when the data obtained from laboratory analysis of that physical sample becomes available to users of this database? - **For the NAMS and SLAMS data, the federal regulations require that the data be supplied within 90 days after the end of the calendar quarter. Most states supply their required data within 60 days. PAMS data also has reporting requirements that can range between 90 days and 6 months depending on the pollutant.**

5. Internal Consistency

"How well does this database support comparisons of environmental conditions in different places or at a single place at different points in time?"

a. Degree of National Consistency in the Administration and Management of this Data Collection Activity

(1) Degree of centralization of key administrative functions

(a) Is the administration of the Data Collection Design centralized or decentralized? - **Centralized for the National networks.**
(b) Is the administration of the Data Collection itself centralized or decentralized? - **Decentralized**
c) Is the administration of the Data Collection editing centralized or decentralized? - **Centralized within the AIRS system**
(d) Is the administration of the Data Collection summarization/estimation centralized or decentralized? - **Centralized within the AIRS system**
(e) Is the administration of the publication strategy for this Data Collection Activity centralized or decentralized? **Centralized if you are speaking of the National Trends Report**

(2) Total program office resources devoted to this data collection activity
(a) Estimated Number of Full Time Staff in Data Collection Activity _0_ -*****
(b) Estimated Amount of Contract Funds for this Data Collection Activity $__0__K -*

(3) Number of full time program office staff carrying out the key administrative functions
(a) Number of full time staff supporting Data Collection Design. **0***
(b) Number of full time staff supporting Data Collection itself. **0***
(c) Number of full time staff supporting Data Collection editing. **1.5**
(d) Number of full time staff supporting Data Collection summarization/estimation. **0***
(e) Number of full time staff supporting publication of the Data Collection Activity. **0***

*** - OAQPS is not involved with the actual collection of the data. OAQPS' involvement is only in the storage and analysis of the data. Please note that there is a huge amount of effort expended by the state and local agencies to collect this data. The number of hours used to collect the required ambient information for EPA has been estimated at 1.8 million hours per year.**

(4) Presence of other operational elements that work to promote national consistency

(a) Do you have a yearly conference on this Data Collection Activity? - **Yes**
(b) Do this Data Collection Activity have a "guidance document" describing the overall administration of this Data Collection Activity? - **Yes**
(c) Is the process of developing and writing the "guidance document" administered centrally? - **Yes**
(d) Is the raw data publicly available on the Internet? -**Not Yet...July, 1997** HTTP Address:
(e) Is the summarized data publicly available on the Internet?-**Yes** HTTP Address:**www.epa.gov/airs/aexec.html**

(5) Use of written manuals/documentation for key elements of this data collection activity? In particular:

(a) Do you have written Data Collection Design Manuals/Documentation? - **Yes** Y/N
(b) Do you have written Data Collection Manuals/Documentation? - **Yes** Y/N
(c) Do you have written Data Editing and validation Manuals/Documentation? - **Yes**
(d) Do you have written Data Summarization/Estimation Manuals/Documentation? - **Yes** Y/N
(e) Do you have a written publication strategy for this Data Collection Activity? Y/N**unknown**

(6) Use of training to support key elements of this data collection activity

(a) Do you provide training in Data Collection Design? - **Yes** Y/N
(b) Do you provide training in the Data Collection process? - **Yes** Y/N
(c) Do you provide training in the Data Editing? - **Yes** Y/N
(d) Do you provide training in the Data Summarization/Estimation? - **Yes** Y/N
(e) Do you provide training on publication strategy for this Data Collection Activity? - **Not to my knowledge** Y/N

c. Use of QA/QC to promote consistency

(1) Are there documented QA/QC procedures for this data collection activity?

(a) Are the QA/QC procedures in conformance with EPA QA/QC guidance and requirements? - **Yes**

(b) Are these procedures actually being applied as specified? - **Yes**

c) Do these procedures include data quality audits? - **Yes**

(I) If so, what percentage of the data generated have been audited?
5%
(II) If so, what have these data quality audits shown for the following with regard to the following:

aa. Missing data

1. To what extent is there missing data in the data set as a whole and at specific data collection sites? -**The only indicator that we have to show how much data is missing is by establishing system criteria that the data is "valid" if they have a data capture rate of 75%. Most of the data in the system meets this criteria.**

2. Are there any observed patterns to what data is missing that a person making use of the data should be aware of? - **No**

bb. Gross errors

1. What is the overall error rate for gross errors (e.g., misplaced decimal point, corrupted data or site information, etc.)**Much less than 0.01%**

2. What did the audit reveal to be the sources of these gross errors? What percentage of the gross errors were due to each source? What kind of error did each source result in? (E.g., shift of the decimal point one figure to the right, switching of data for two different parameters [e.g., the DO value was entered in pace of the suspected solids value and the suspected solids value was entered for the DO value], etc.)
- **Miscoding of Null data reason codes;**

cc. Precision and accuracy

1. What did the audit reveal to be the overall precision and accuracy of the data in the database (after excluding gross errors)?
Precision: -7.0% to +7.5%
Accuracy (Level 1): -6.9% to +6.0%
Accuracy (Level 2): -5.6% to +4.4%

2. Did these vary from site to site, seasonally, diurnally, or through time? - **Yes - Yes - Yes - Yes**

a. If so, in what ways did they vary? - **in all the forms described**

b. If so, what was determined to be the cause of this variation? - **No determination documented**

d. Consistency in Comprehensiveness, Spatial Characteristics and Temporal Characteristics

(1) Consistency in "Comprehensiveness"

(a) To what extent is there consistency from site to site and through time in matters addressed in the previous section on "Comprehensiveness" (Section 2, above)? - **75% to 80% of the required data is reported to the system in a timely matter. There is substantial variability from state to state on the number of compounds that are reported.**

(2) Consistency in "Spatial Characteristics"

(b) To what extent is there consistency from site to site and through time in matters addressed in the previous section on "Spatial Characteristics" (Section 3, above)? - **All sites and monitors are under the same rules for the reporting of lat/long coordinates and to specify the accuracy of those coordinates.**

(3) Consistency in "Temporal Characteristics"

c) To what extent is there consistency from site to site and through time in matters addressed in the previous section on "Temporal Characteristics" (Section 4, above)? - **The volume of data through time is highly variable from state to state especially when one considers the "older" data (pre-1980). Since 1979, however, there has been consistency in the reporting of criteria pollutants to the system from all states.**

6. Comparability with and Ability to Integrate with other Databases

"What is the potential for integration of data from this database with data from other databases to prepare a broader picture of the state of the environment than can be obtained from any one database alone?"

a. Physical Integration:

(1) The comparability of this database with other databases and the ability to integrate it with other databases is a determination that must be made on a case-by-case basis. This can be done by comparing the characteristics of this database (as reflected to the responses to the questions in Sections 1-5 above) with the corresponding characteristics of each other database with which there is an interest in integrating it physically. A convenient and useful way to make this comparison is to take the responses (to the questions in sections 1-5 above) for each database to be compared, and to place these responses in a summary matrix of the form shown in Attachment 1.

APPENDIX B

Statistical Profile of EPA databases

How comprehensive is the database?

- Missing data for key variables will determine if the data is truly comprehensive.

- Distribution of the data over key variables in the database.

  For example, AIRS has data on criteria pollutants. How much of the data is devoted to each pollutant? TRI data is reported for facilities of certain specified SIC codes. How many records pertain to each of the SIC codes?

- A frequency distribution of the number of facilities reporting on different numbers of releases or numbers of permits, etc.

  For example, the frequency distribution of the number of different chemicals that TRI facilities report on might be uniform or tightly concentrated on a small number with long tails in the distribution.

Can the data be used for geographic analysis?

- Non-missing data for lat/long, ZIP code, county, and State will indicate the suitability for this purpose (univariate analysis)

  For example, AIRS ,TRI, RCRIS are all required to have data on these spatial variables. It is important to confirm the extent to which the data exist.

- Distribution of the data over some of the geographical scales will provide the user an idea of the extent of data availability. As a univariate analysis, this can be done as a quartile or a percentile analysis to examine if the bulk of the data applies to a few counties or States etc.

  For example, there are six criteria pollutants in AIRS. Is the data coverage skewed towards some States, etc.?

- A multivariate analysis of the distribution indicates the level of data coverage for different pollutants in different geographical areas. This information can be displayed as tables or graphs. This would provide a user information on the relative amount of data available for different identification variables. This analysis can be done for a region, county, State or ZIP code scale.

For example, the State of Montana has X number of records or monitors for sulfur dioxide, Y number for ozone and Z number for NOx. (We have to make sure that the numbers are comparable.) Similarly, a frequency distribution can be done on the number of States with no monitors for a certain pollutant as well as defined ranges of numbers of monitors.

Does this database allow trends analysis?

- Non-missing data will indicate the suitability for this purpose (univariate analysis)

  For example, it is important to confirm the extent to which the data exist for different years.

- Analysis across several years of a database needs to be done, such as a random sampling of data to identify certain States/counties/facilities and track the trends in their information.

  For example, county A had X number of data records in 1990, Y in 1991, and Z in 1992. This can be done for both univariate and multivariate analysis.

How consistent is the database over space and time?

- This can be answered from the above analysis.

Can the data be linked with other databases?

- Non-missing data for spatial variables will indicate this.
- Distribution of data with ID numbers for other databases will also indicate the suitability for linking.
- Picking a random sample of facilities or counties and identifying information from each assessed database will also indicate suitability for linkage.

Are there any checks possible for accuracy?

- If CEIS has lookup software to check on the accuracy of ZIP codes with State and county information, it could check this for spatial analysis.

APPENDIX C

Preliminary Questions for Case Studies

1) What types of environmental information and data does EPA have?

2) Is air and water quality (nationally, regionally, in my community) getting better or worse than it was 5 years ago?

3) What pollutants are being released in my community?

4) What ecosystem do I live in?

5) Is my drinking water safe?

6) What are the impacts of (a plant or facility) on my community?

7) What are the sources of pollution in my community?

8) What are our national and State air and water standards? Is my community meeting these standards?

9) If I'm living in an area where these standards are not being met, what does that mean in terms of public health?

10) Where can I get more detailed information?

11) Are the fish in the nearby river or lake safe to eat?

12) How does my community compare with others in the country?

13) Are there any companies or businesses in my community who are out of compliance with your regulations?

14) You report that my community is out of attainment for nitrogen oxides (or other pollutant). How much of that problem comes from "x" or "y" sources (e.g., point, non-point, cars, businesses, public facilities, etc.)?

15) I have a problem with the way that you have listed my company as a source of pollution. Who can I talk to, in order to fix this problem?

16) Can someone explain how to use the data and information that you are presenting on this website?

17) You report that "x" percent of the rivers in the country are not meeting their designated uses. What's a designated use?

18) What toxics am I being exposed to in my community?

19) What are the health effects of "x" (name of a specific pollutant or contaminant)?

20) What are the sources of greenhouse gases in my community and what can we do to control them?